MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION F

AD-A264 015

-0188

(2)

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3.
	February 1993	Technical

4. TITLE AND SUBTITLE
Easy-to-Apply Results for Establishing Convergence of Markov Chains in Bayesian Analysis

5. FUNDING NUMBERS
DAAL03-90-G-0103

6. AUTHOR(S)
Krishna B. Athreya, Hani Doss and Jayaram Sethuraman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Statistics
Florida State University
Tallahassee, FL 32306-3033

8. PERFORMING ORGANIZATION REPORT NUMBER
FSU Technical Report No. 884

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 27868.26-MA

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Abstract

The Markov chain simulation method has become a powerful computational method in Bayesian analysis. The success of this method depends on the convergence of the Markov chain to its stationary distribution. We give two carefully stated theorems, whose conditions are easy to verify, that establish this convergence. We give versions of our conditions which are simpler to verify for the Markov chains that arise most commonly in Bayesian analysis.

14. SUBJECT TERMS
Bayesian Poisson regression; calculation of posterior districutions; ergodic theorem; Markov chain simulation method.

15. NUMBER OF PAGES
7

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Easy-to-Apply Results for Establishing Convergence of Markov Chains in Bayesian Analysis

Krishna B. Athreya[1]

Department of Statistics
Iowa State University
Ames, Iowa 50011

Hani Doss[2] and Jayaram Sethuraman[3]

Department of Statistics
Florida State University
Tallahassee, Florida 32306–3033

February 1993

FSU Technical Report No. 884
AFOSR Series D Technical Report No. 10
USARO Technical Report No. 133

---

[1] Research supported by National Science Foundation Grant DMS-92-04938.

[2] Research supported by Air Force Office of Scientific Research Grant 90-0202.

[3] Research supported by Army Research Office Grant DAAL03-90-G-0103.

Easy-to-Apply Results for Establishing Convergence of Markov Chains in Bayesian Analysis

Krishna B. Athreya *
Iowa State University

Hani Doss †and Jayaram Sethuraman ‡
Florida State University

Abstract

The Markov chain simulation method has become a powerful computational method in Bayesian analysis. The success of this method depends on the convergence of the Markov chain to its stationary distribution. We give two carefully stated theorems, whose conditions are easy to verify, that establish this convergence. We give versions of our conditions which are simpler to verify for the Markov chains that arise most commonly in Bayesian analysis.

Key words and phrases: Bayesian Poisson regression; calculation of posterior distributions; ergodic theorem; Markov chain simulation method.

1 Introduction

Let π be a probability distribution on a measurable space $(\mathcal{X}, \mathcal{B})$. The Monte Carlo Markov chain method is a technique for estimating characteristics of π such as $\pi(E)$ or $\int f d\pi$ where $E \in \mathcal{B}$ and f is a bounded measurable function, and which is useful when π is too complex to describe analytically. The idea is very

*Research supported by National Science Foundation Grant DMS-92-04938
†Research supported by Air Force Office of Scientific Research Grant 90-0202
‡Research supported by Army Research Office Grant DAAL03-90-G-0103

straightforward. We construct a transition probability function $P(x, \cdot)$ with the property that it has stationary distribution π, i.e.

$$\pi(C) = \int P(x, C)\pi(dx) \quad \text{for all } C \in \mathcal{B}. \tag{1.1}$$

Then, we generate a Markov chain $\{X_n\}$ with this transition probability function as follows. We fix a starting point x_0, generate an observation X_1 from $P(x_0, \cdot)$, generate an observation X_2 from $P(X_1, \cdot)$, etc. This produces the Markov chain $x_0 = X_0, X_1, X_2, \ldots$. We use this construction in one of two ways. Either we discard an initial segment $X_0, X_1, X_2, \ldots, X_r$ of the Markov chain, in which the chain has not yet converged to its stationary distribution, and retain the rest of the chain, or we independently run a large number of chains and for each retain only the last observation. In either case we we use the observations that we have kept to obtain empirical estimates of those features of π that are of interest.

Implicit in this method is the assumption that the chain converges to its stationary distribution, for a wide class of starting points x_0. Indeed, one can easily give examples of Markov chains that do not converge to their stationary distribution from any starting point. Thus, to establish the validity of the method, it is crucial to obtain results that give conditions which imply convergence of the chain.

The Markov chain literature already contains many results that give conditions under which the Markov chain converges to its stationary distribution for a class of starting points x_0 which have probability one under π (this condition is called ergodicity). Unfortunately, when one comes to apply these results, one immediately notices that in *statistical* applications, the conditions of these theorems are virtually impossible to check.

In our work we have obtained two theorems (Theorems 1 and 2 below) that assert ergodicity of the chain under conditions that are extremely easy to verify in a wide range of problems that are likely to arise in Bayesian statistics. These theorems pertain, roughly, to the two modes of using the Markov chain construction. Before explaining our theorems, it is useful to give an idea of the wide scope of the problems that can be approached via the Monte Carlo Markov chain method.

There are many ways to produce a transition function satisfying (1.1). Methods include the Metropolis algorithm and its variants, and the so-called Gibbs sampler. This last method appears to be the one that is the most widely used in Bayesian statistics, and we now proceed to describe it. This algorithm is used to estimate the unknown joint distribution $\pi = \pi_{X^{(1)},\ldots,X^{(p)}}$ of the (possibly vector-valued) random variables $(X^{(1)}, \ldots, X^{(p)})$ by updating the coordinates one at a time, as follows. We suppose that we know the conditional distributions $\pi_{X^{(i)}|\{X^{(j)}\, j\neq i\}}$, $i = 1, \ldots, p$ or at least that we are able to generate observations from these conditional distributions. If $X_m = (X_m^{(1)}, \ldots, X_m^{(p)})$ is the current state, the next state $X_{m+1} = (X_{m+1}^{(1)}, \ldots, X_{m+1}^{(p)})$ of the Markov chain is formed as follows. Generate $X_{m+1}^{(1)}$ from $\pi_{X^{(1)}|\{X^{(j)}\, j\neq 1\}}(\cdot, X_m^{(2)}, \ldots, X_m^{(p)})$, then $X_{m+1}^{(2)}$ from

$\pi_{X^{(2)}|\{X^{(j)}\ j\neq 2\}}(X^{(1)}_{(m+1)}, \cdot, X^{(3)}_m, \ldots, X^{(p)}_m)$, and so on until $X^{(p)}_{m+1}$ is generated from $\pi_{X^{(p)}|\{X^{(j)}\ j\neq p\}}(X^{(1)}_{(m+1)}, \ldots, X^{(p-1)}_{(m+1)}, \cdot)$. If P is the transition function that produces X_{m+1} from X_m, then it is easy to see that P satisfies (1.1).

We now give a very brief description of how this method is useful in some Bayesian problems. We suppose that the parameter θ has some prior distribution, that we observe a data point Y whose conditional distribution given θ is $\mathcal{L}(Y \mid \theta)$, and that we wish to obtain $\mathcal{L}(\theta \mid Y)$, the conditional distribution of θ given Y. It is often the case that if we consider an (unobservable) auxiliary random variable Z, then the distribution $\pi_{\theta,Z} = \mathcal{L}(\theta, Z \mid Y)$ has the property that $\pi_{\theta|Z}$ $(= \mathcal{L}(\theta \mid Y, Z))$ and $\pi_{Z|\theta}$ $(= \mathcal{L}(Z \mid Y, \theta))$ are easy to calculate. Typical examples are missing and censored data problems. If we have a conjugate family of prior distributions on θ, then we may take Z to be the missing or the censored observations, so that $\pi_{\theta|Z}$ is easy to calculate. The Gibbs sampler then gives a random observation with distribution (approximately) $\mathcal{L}(\theta, Z \mid Y)$, and retaining the first coordinate gives an observation with distribution (approximately) equal to $\mathcal{L}(\theta \mid Y)$.

Another application arises when the parameter θ is high dimensional, and we are in a nonconjugate situation. Let us write $\theta = (\theta_1, \ldots, \theta_k)$, so that what we wish to obtain is $\pi_{\theta_1,\ldots,\theta_k}$. Direct calculation of the posterior will involve the evaluation of a k-dimensional integral, which may be difficult to accomplish. On the other hand, application of the Gibbs sampler involves the generation of one-dimensional random variables from $\pi_{\theta_i|\{\theta_j\ j\neq i\}}$. The generation of random variables from a one-dimensional distribution is in general much easier than from a multidimensional distribution; very often special tricks can be used. We illustrate this with an example in Section 2 below. In addition, we note that the distribution $\pi_{\theta_i|\{\theta_j\ j\neq i\}}$ is available in closed form, except for a normalizing constant. There exist very efficient algorithms for generating random variables from such a distribution, provided the distribution is unimodal; see Zaman (1992).

2 Illustration of the Markov Chain Simulation Method: Bayesian Poisson Regression

As a typical application of how the Gibbs sampler helps in high dimensional problems, we consider a model involving Bayesian Poisson regression. This model is

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = \textstyle\sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, 2, \ldots, n,$$

where the x_{ij}'s are non-negative covariates, and where the prior distribution on the β_j's is a product of Gammas. In this case, the likelihood function is

$$p(\boldsymbol{\lambda}) \;=\; \prod_{i=1}^{n} \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!}$$

$$= (y_1!\, y_2! \ldots y_n!)^{-1} \exp\Big(-\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\beta_j\Big) \prod_{i=1}^{n}\Big(\sum_{j=1}^{p} x_{ij}\beta_j\Big)^{y_i}$$

and the joint density of the β_j's is given by

$$\begin{aligned} f_{\beta}(\beta_1, \beta_2, \ldots, \beta_p) &= \prod_{j=1}^{p} \frac{b_j^{a_j}}{\Gamma(a_j)} \beta_j^{a_j - 1} \exp(-b_j \beta_j) \\ &\propto \exp\Big(-\sum_{j=1}^{p} b_j \beta_j\Big) \prod_{j=1}^{p} \beta_j^{a_j - 1}, \end{aligned}$$

where a_j is the shape and b_j is the scale parameter for the distribution of β_j, $j = 1, 2, \ldots, p$. The posterior joint density of the β_j's, given the data, is therefore

$$\pi(\beta_1, \beta_2, \ldots, \beta_p) \propto \exp\Big(-\sum_{j=1}^{p} \beta_j v_j\Big) \prod_{j=1}^{p} \beta_j^{a_j - 1} \Big(\prod_{i=1}^{n} (\sum_{j=1}^{p} x_{ij}\beta_j)^{y_i}\Big).$$

where $v_j = b_j + \sum_{i=1}^{n} x_{ij}$, $j = 1, 2, \ldots, p$. To determine the posterior joint density of the β_j's exactly, the constant of proportionality needs to be determined. This requires high-dimensional integration. However, the Gibbs sampler can be used if we know the conditional distributions of any β_i given the rest of the β_j's and the data.

To compute the conditional density of any β_k, $k = 1, 2, \ldots, p$, given the rest of the β_j's and the data, we proceed as follow. For each l, $1 \le l \le p$, let $S_l = \{1, 2, \ldots, p\} \setminus \{l\}$. Then for each k, the density of β_k, conditional on all β_j, $j \in S_k$, and the data is the discrete mixture of Gamma densities

$$f_{\beta_k | \beta_j, j \in S_k}(\beta_k) \propto \beta_k^{a_k - 1} \exp(-v_k \beta_k) \prod_{i=1}^{n} (c_i + x_{ik}\beta_k)^{y_i},$$

where $c_i = \sum_{j \in S_k} x_{ij}\beta_j$. Let $m = \sum_{i=1}^{n} y_i$ and write

$$\prod_{i=1}^{n} (c_i + x_{ik}\beta_k)^{y_i} = \sum_{l=0}^{m} r_l(k) \beta_k^l,$$

where we explicitly show the dependence of the coefficients on k. Then,

$$f_{\beta_k | \beta_j, j \in S_k}(\beta_k) \propto \sum_{l=0}^{m} r_l(k) \beta_k^{a_k + l - 1} \exp(-v_k \beta_k)$$

and we readily recognize that

$$f_{\beta_k | \beta_j, j \in S_k}(\beta_k) = \sum_{l=0}^{m} p_l(k) g_{a_k + l, v_k}(\beta_k),$$

where $g_{p,q}(x)$ denotes the gamma density with shape parameter p and scale parameter q in x, and $p_l(k) = r_l(k)\Gamma(a_k + l)/v_k^{a_k + l}$. The $p_l(k)$'s are the discrete mixture probabilities.

3 Convergence Theorems

Before stating our theorems, we will need a few definitions concerning Markov chains. Let $P^n(x,\cdot)$ denote the distribution of X_n when the chain is started at x. Also, for a set $C \in \mathcal{B}$, let $T(C) = \inf\{n : n > 0,\ X_n \in C\}$ be the first time the chain hits C, after time 0. Finally, for any subset $\mathcal{I}$ of the positive integers, g.c.d.$(\mathcal{I})$ will denote the greatest common divisor of the integers in $\mathcal{I}$.

Theorem 1 *Suppose that the Markov chain $\{X_n\}$ with transition function $P(x,C)$ has an invariant probability measure π, i.e. (1.1) holds. Suppose that there is a set $A \in \mathcal{B}$, a probability measure ρ with $\rho(A) = 1$, a constant $\epsilon > 0$, and an integer $n_0 \geq 1$ such that*

$$\pi\Big\{x : P_x(T(A) < \infty) > 0\Big\} = 1, \tag{3.1}$$

and

$$P^{n_0}(x,\cdot) \geq \epsilon\rho(\cdot) \quad \text{for each } x \in A. \tag{3.2}$$

Suppose further that

$$\text{g.c.d.}\Big\{m \geq 1 : \text{ there is an } \epsilon_m > 0 \text{ such that } \sup_{x\in A} P^m(x,\cdot) \geq \epsilon_m\rho(\cdot)\Big\} = 1. \tag{3.3}$$

Then there is a set D_0 such that

$$\pi(D_0) = 1 \quad \text{and} \quad \sup_{C\in\mathcal{B}} |P^n(x,C) - \pi(C)| \to 0 \quad \text{for each } x \in D_0. \tag{3.4}$$

Theorem 2 *Suppose that the Markov chain $\{X_n\}$ with transition function $P(x,C)$ satisfies conditions (1.1), (3.1) and (3.2). Then*

$$\sup_{C\in\mathcal{B}} \Big|\frac{1}{n_0}\sum_{r=0}^{n_0-1} P^{mn_0+r}(x,C) - \pi(C)\Big| \to 0 \quad \text{as } m \to \infty \quad \text{for } [\pi]\text{-almost all } x, \tag{3.5}$$

and hence

$$\sup_{C\in\mathcal{B}} \Big|\frac{1}{n}\sum_{j=1}^{n} P^j(x,C) - \pi(C)\Big| \to 0 \quad \text{as } n \to \infty \quad \text{for } [\pi]\text{-almost all } x. \tag{3.6}$$

Let $f(x)$ be a measurable function on $(\mathcal{X},\mathcal{B})$ such that $\int \pi(dy)|f(y)| < \infty$. Then

$$P_x\Big\{\frac{1}{n}\sum_{j=1}^{n} f(X_j) \to \int \pi(dy)f(y)\Big\} = 1 \quad \text{for } [\pi]\text{-almost all } x \tag{3.7}$$

and

$$\frac{1}{n}\sum_{j=1}^{n} E_x(f(X_j)) \to \int \pi(dy)f(y) = 1 \quad \text{for } [\pi]\text{-almost all } x. \tag{3.8}$$

Theorem 1 requires condition (3.3), while Theorem 2 does not. Theorem 2 states that if condition (3.3) is violated, one can still apply the Markov chain simulation method, except that one has to work with averages of dependent random variables instead of running a large number of independent chains and working with an (approximately) independent sample. These two theorems are proved in Athreya, Doss, and Sethuraman (1992), where it is also shown that these are the weakest possible conditions that will ensure convergence of a Markov chain for a set of starting points having probability one under the stationary distribution.

There are already many theorems that give conditions that guarantee ergodicity of Markov chains. See the discussion in Section 1 of Athreya, Doss, and Sethuraman (1992). Most of these theorems are stated under two general classes of conditions. Conditions in the first class involve verification of a "recurrence condition" which is much stronger than our condition (3.1). Conditions in the second class of involve the stationary distribution of the chain. Since this stationary distribution is unknown, these conditions are difficult to verify. In contrast, our theorems are stated under conditions that involve only the transition function, and thus are, in general, easier to verify.

Theorems 1 and 2 pertain to arbitrary Markov chains. As we mentioned earlier, the Gibbs sampler is the most commonly used Markov chain in Bayesian statistics. We now give a result that facilitates the use of our theorems when the Markov chain used is the Gibbs sampler. We use the notation of Section 1, and assume that for each i, the conditional distributions $\pi_{X_i|\{X^{(j)}\, j\neq i\}}$ have densities, say $p_{X_i|\{X^{(j)}\, j\neq i\}}$, with respect to some dominating measure ρ_i.

Theorem 3 *Suppose that for each $i = 1, \ldots, k$ there is a set A_i with $\rho_i(A_i) > 0$, and a $\delta > 0$ such that for each $i = 1, \ldots, k$*

$$p_{X_i|\{X^{(j)}\, j\neq i\}}(x^{(1)}, \ldots, x^{(k)}) > 0 \tag{3.9}$$

whenever

$$x^{(1)} \in A_1, \ldots, x^{(i)} \in A_i, \text{ and } x^{(i+1)}, \ldots, x^{(k)} \text{ are arbitrary,}$$

and

$$p_{X_i|\{X^{(j)}\, j\neq i\}}(x^{(1)}, \ldots, x^{(k)}) > \delta \text{ whenever } x^{(j)} \in A_j,\ j = 1, \ldots, k.$$

Then conditions (3.1) and (3.2) are satisfied with $n_0 = 1$. Thus, (3.3) is also satisfied, and the conclusions of Theorems 1 and 2 hold.

We note that condition (3.9) is often checked for all $x^{(1)}, \ldots, x^{(k)}$.

This theorem is immediate for the case $k = 2$. For the general case the proof follows by induction.

References

Athreya, K. B., Doss, H., and Sethuraman, J. (1992). A proof of convergence of the Markov chain simulation method. Technical Report No. 868, Department of Statistics, Florida State University.

Doss, H. and Narasimhan, B. (1993). Bayesian Poisson regression using the Gibbs sampler: A case study in dynamic graphics. Technical Report, Department of Statistics, Florida State University.

Zaman, A. (1992). Generating random numbers from a unimodal density by cutting corners. Technical Report, Department of Statistics, Florida State University.